

# Rethinking the Reference model in RLHF

YiMing Liu, letusgo126@126.com

March 2025

## 1 贡献

1. 首次提供了分类损失中 KL 的梯度无偏估计，证明了 k1 和 k2 的转换关系，k2 估计作为 KL loss 比 k3 好 (k3 的梯度是 k2 的有偏估计)，提供了 PPO 原始 reward 公式的推导。
2. 首次提供了 Diffusion RLHF 的 DDPO 如何加入 KL 惩罚项，解决了 Diffusion RLHF 中必须额外准备数据集加入 pretrain loss 的问题，降低了额外数据集带来分布偏移的影响。
3. 首次证明了目前 DiffusionDPO 推导的错误之处。

## 2 基于策略梯度方法的 PPO 奖励函数推导

### 2.1 RLHF 目标函数建模

基于人类反馈的强化学习 (RLHF) 的核心目标可建模为以下优化问题：

$$\arg \max_{\theta} J = \underbrace{\mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_{\theta}(y|x)} [r(x, y)]}_{\text{奖励最大化项}} - \beta \underbrace{\mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{\pi_{\theta}(\cdot|s_t, x)} [\mathbb{D}_{\text{KL}}(\pi_{\theta} || \pi_{\text{ref}})]}_{\text{策略约束项}} \quad (1)$$

其中  $\beta$  为温度系数，控制策略偏离参考模型  $\pi_{\text{ref}}$  的程度。该目标函数包含两个关键部分：奖励期望的最大化和 KL 散度约束的平衡。

### 2.2 目标函数分解与分析

首先考虑奖励相关项的损失函数构造。根据策略梯度定理，可建立奖励项的损失函数为：

$$-\mathcal{L}_R(\theta) = \mathbb{E}_{x \sim \mathcal{D}, a_t \sim \pi_{\theta}(a_t|s_t, x)} [r(x, y) \log \pi_{\theta}(a_t|s_t, x)] \quad (2)$$

结合 KL 散度约束项  $\mathcal{L}_{\text{KL}_t}(\theta, \text{ref})$ ，完整损失函数可表示为：

$$L_{\text{total}} = -\arg \min_{\theta} J = -(\mathcal{L}_R(\theta) - \beta \mathcal{L}_{\text{KL}_t}(\pi_{\theta}, \pi_{\text{ref}})) \quad (3)$$

### 2.3 梯度推导过程

对目标函数进行梯度分析时，需要分别处理两个组成部分：

1. 奖励项梯度：

$$\nabla_{\theta} \mathcal{L}_R = \mathbb{E}_{x \sim \mathcal{D}, a_t \sim \pi_{\theta}(a_t | s_t, x)} [r(x, y) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t, x)] \quad (4)$$

2. KL 散度项梯度（具体推导见引理 eq. (47)）：

$$-\nabla_{\theta} \mathcal{L}_{\text{KL}_t} = -\mathbb{E}_{x \sim \mathcal{D}, a_t \sim \pi_{\theta}(a_t | s_t, x)} \left[ \log \frac{\pi_{\theta}(a_t | s_t, x)}{\pi_{\text{ref}}(a_t | s_t, x)} \cdot \nabla_{\theta} \log \pi_{\theta}(a_t | s_t, x) \right] \quad (5)$$

将二者结合可得总梯度：

$$\begin{aligned} \nabla_{\theta} J &= \nabla_{\theta} \mathcal{L}_R - \beta \nabla_{\theta} \mathcal{L}_{\text{KL}_t} \\ &= \mathbb{E}_{x, a_t \sim \pi_{\theta}(a_t | s_t, x)} \left[ \left( r(x, y) - \beta \log \frac{\pi_{\theta}(a_t | s_t, x)}{\pi_{\text{ref}}(a_t | s_t, x)} \right) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t, x) \right] \end{aligned} \quad (6)$$

### 2.4 等效奖励函数构造

通过梯度分析可发现，原始优化问题可等价转换为：

$$\arg \max_{\theta} J' = \mathbb{E}_{x \sim \mathcal{D}, a_t \sim \pi_{\theta}} \left[ r(x, y) - \beta \log \frac{\pi_{\theta}(a_t | s_t, x)}{\pi_{\text{ref}}(a_t | s_t, x)} \right] \quad (7)$$

这揭示了 PPO 算法中奖励函数的设计本质：

$$r_t = r(x, y) - \beta \log \frac{\pi_{\theta}(a_t | s_t, x)}{\pi_{\text{ref}}(a_t | s_t, x)} \quad (8)$$

### 2.5 关键前提条件

需要特别强调的是，上述推导成立的关键在于 KL 散度项的梯度必须满足 K1 及其类似的特定形式。当且仅当形式满足 k1 或类似形式时才成立。若该近似条件不成立，则公式 eq. (7) 和 eq. (8) 的推导过程将失效，因为 DPO 的推导基础 eq. (65) 就是该公式。

### 3 KL 惩罚项的数学性质分析

#### 3.1 奖励函数中 KL 惩罚项的适用性条件

在 PPO 算法设计中，仅特定形式的 KL 散度估计量适合作为奖励函数的惩罚项。根据理论分析，k1 估计量具有数学适用性，其表达式如 eq. (8) 所示。然而，k2 和 k3 估计量由于内在的数学性质缺陷，不适合作为惩罚项，原因如下：

考虑 k2 和 k3 的估计值  $kl_t$  具有恒正特性，其梯度方向始终满足：

$$-\nabla_{\theta} \text{KL}_t(\pi_{\theta} \parallel \pi_{ref}) = -kl_t \nabla_{\theta} \log \pi_{\theta}(a_t | s_t, x) \quad (9)$$

此时无论  $\pi_{\theta}(a_t | s_t, x)$  与  $\pi_{ref}(a_t | s_t, x)$  的概率分布关系如何，梯度更新方向都会强制降低当前策略  $\pi_{\theta}$  生成任意动作  $a_t$  的概率。这种单向的惩罚机制本质上违背了策略优化算法的基本设计原则，即应建立方向可调的奖惩机制来引导策略改进。

#### 3.2 KL 惩罚项的损失函数适配性

**k2 估计量的有效性** 如 eq. (48) 所示，k2 估计量能够构造有效的 KL 惩罚损失函数。其数学本质是通过 KL 散度的对称性设计，建立双向调节机制：当  $\pi_{\theta}$  偏离  $\pi_{ref}$  时，梯度方向会根据偏离方向自动调整，既防止策略过度偏离，又保留必要的优化自由度。

**k1 估计量的失效机理** k1 对应的损失函数形式为：

$$\mathcal{L}_{\text{KL}_t}(\pi_{\theta}, \pi_{ref}) = \log \pi_{\theta}(a_t | s_t, x) - \log \pi_{ref}(a_t | s_t, x) \quad (10)$$

其梯度表达式揭示本质缺陷：

$$-\nabla_{\theta} \mathcal{L}_{\text{KL}_t} = -\nabla_{\theta} \log \pi_{\theta}(a_t | s_t, x) \quad (11)$$

该梯度恒指向降低当前策略概率的方向，形成类似极大似然估计的反向约束。这种单向作用机制会导致策略网络的概率输出持续衰减，最终引发模型坍缩问题。

**k3 估计量的近似特性** k3 构造的损失函数具有特殊形式：

$$\mathcal{L}_{\text{KL}_t} = \frac{\pi_{\text{ref}}(a_t|s_t, x)}{\pi_\theta(a_t|s_t, x)} - \log \frac{\pi_{\text{ref}}(a_t|s_t, x)}{\pi_\theta(a_t|s_t, x)} - 1 \quad (12)$$

对应的梯度表达式揭示其近似本质：

$$-\nabla_\theta \mathcal{L}_{\text{KL}_t} = \left( \frac{\pi_{\text{ref}}}{\pi_\theta} - 1 \right) \nabla_\theta \log \pi_\theta(a_t|s_t, x) \quad (13)$$

令  $x = \pi_{\text{ref}}/\pi_\theta$ ，k2 与 k3 的梯度可对比表示为：

$$\text{k2 梯度} : \log x \cdot \nabla_\theta \log \pi_\theta \quad (14)$$

$$\text{k3 梯度} : (x - 1) \cdot \nabla_\theta \log \pi_\theta \quad (15)$$

在策略邻域 ( $x \approx 1$ ) 进行泰勒展开时,  $\log x \approx x - 1$ , 此时 k3 梯度构成 k2 梯度的线性近似。但这种近似具有两个关键缺陷：1. **有偏性**：当策略显著偏离参考策略 ( $x$  远离 1, 也就是训练后期  $\pi_{\text{ref}}$  离  $\pi_\theta$  较远) 时, 近似误差呈非线性增长 2. **非对称性**： $x - 1$  对  $\pi_\theta > \pi_{\text{ref}}$  和  $\pi_\theta < \pi_{\text{ref}}$  的响应特性不对称。

## 4 Diffusion RLHF 之 DDPO

$\pi$  表示概率, 常用符号为  $p$ 。

### 4.1 扩散模型基础

Denoising Diffusion Probabilistic Models (DDPM) 的逆向过程采样公式定义为：

$$\pi_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c}) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, \mathbf{c}, t), \sigma_t^2 \mathbf{I}) \quad (16)$$

其具体采样形式为：

$$\mathbf{x}_{t-1} = \boldsymbol{\mu}_\theta(\mathbf{x}_t, \mathbf{c}, t) + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (17)$$

其中  $\mathbf{c}$  为条件信息,  $\sigma_t^2$  为预定义的方差参数。

## 4.2 概率密度与梯度推导

当方差固定时，高维高斯分布的概率密度函数可展开为：

$$\pi_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c}) = \frac{1}{(2\pi\sigma_t^2)^{d/2}} \exp\left(-\frac{\|\mathbf{x}_{t-1} - \boldsymbol{\mu}_{\theta}(\mathbf{x}_t, \mathbf{c}, t)\|^2}{2\sigma_t^2}\right) \quad (18)$$

其中  $d$  为潜变量维度。取对数概率得：

$$\log \pi_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c}) = -\frac{d}{2} \ln(2\pi) - \frac{d}{2} \ln \sigma_t^2 - \frac{\|\mathbf{x}_{t-1} - \boldsymbol{\mu}_{\theta}(\mathbf{x}_t, \mathbf{c}, t)\|^2}{2\sigma_t^2} \quad (19)$$

对模型参数  $\theta$  求梯度时，前两项为常数项，故梯度表达式简化为：

$$\nabla_{\theta} \log \pi_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c}) = -\frac{1}{2\sigma_t^2} \nabla_{\theta} \|\mathbf{x}_{t-1} - \boldsymbol{\mu}_{\theta}(\mathbf{x}_t, \mathbf{c}, t)\|^2 \quad (20)$$

## 4.3 Diffusion 强化学习目标函数

DDPO (Diffusion Policy Optimization) 的优化目标定义为：

$$\arg \max_{\theta} J = \underbrace{\mathbb{E}_{\mathbf{c} \sim \mathcal{D}, \mathbf{x}_0 \sim \pi_{\theta}(\cdot | \cdot, \mathbf{c})} [r(\mathbf{x}_0, \mathbf{c})]}_{\text{奖励最大化项}} \quad (21)$$

DDPO 的梯度函数定义为：

$$\nabla_{\theta} J = \nabla_{\theta} \mathcal{L}_{DDPO} = \mathbb{E}_{\mathbf{c} \sim \mathcal{D}, t \sim U(0, T)} [r(\mathbf{x}_0, \mathbf{c}) \nabla_{\theta} \log \pi_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c})] \quad (22)$$

展开后得到完整表达式：

$$\begin{aligned} -\mathcal{L}_{DDPO} &= \mathbb{E}_{\mathbf{c} \sim \mathcal{D}, t \sim U(0, T)} [r(\mathbf{x}_0, \mathbf{c}) \log \pi_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c})] \\ &= \mathbb{E}_{\mathbf{c} \sim \mathcal{D}, t \sim U(0, T)} \left[ r(\mathbf{x}_0, \mathbf{c}) \left( \underbrace{-\frac{d}{2} \ln(2\pi) - \frac{d}{2} \ln \sigma_t^2}_{\text{常数项}} - \frac{\|\mathbf{x}_{t-1} - \boldsymbol{\mu}_{\theta}\|^2}{2\sigma_t^2} \right) \right] \end{aligned} \quad (23)$$

注意到常数项对优化无贡献，可得等价目标函数：

$$-\mathcal{L}_{DDPO_{equiv}} = \mathbb{E}_{\mathbf{c} \sim \mathcal{D}, t \sim U(0, T)} \left[ -r(\mathbf{x}_0, \mathbf{c}) \frac{\|\mathbf{x}_{t-1} - \boldsymbol{\mu}_{\theta}(\mathbf{x}_t, \mathbf{c}, t)\|^2}{2\sigma_t^2} \right] \quad (24)$$

## 4.4 与预训练目标的关联

对比扩散模型预训练目标：

$$-\mathcal{L}_{pretrain}(\theta) = \mathbb{E}_{(\mathbf{x}_0, \mathbf{c}) \sim p(\mathbf{x}_0, \mathbf{c}), t \sim \mathcal{U}\{0, T\}, \mathbf{x}_t \sim q(\mathbf{x}_t | \mathbf{x}_0)} \left[ -\frac{\|\tilde{\boldsymbol{\mu}}(\mathbf{x}_0, t) - \boldsymbol{\mu}_{\theta}(\mathbf{x}_t, \mathbf{c}, t)\|^2}{2\sigma_t^2} \right] \quad (25)$$

可知 DDPO 在预训练目标基础上引入了奖励加权机制，实现对齐优化。

## 4.5 重要性采样扩展

引入重要性采样比后，DDPO 改进形式为：

$$\begin{aligned} -\mathcal{L}_{DDPO_{IS}} &= \mathbb{E}_{\mathbf{c} \sim D, t \sim U(0, T)} \left[ r(\mathbf{x}_0, \mathbf{c}) \frac{p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c})}{p_{\theta_{old}}(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c})} \right] \\ &= \mathbb{E}_{\mathbf{c} \sim D, t \sim U(0, T)} \left[ r(\mathbf{x}_0, \mathbf{c}) \exp \left( \frac{\|\mathbf{x}_{t-1} - \boldsymbol{\mu}_{\theta_{old}}(\mathbf{x}_t, \mathbf{c}, t)\|^2}{2\sigma_t^2} - \frac{\|\mathbf{x}_{t-1} - \boldsymbol{\mu}_{\theta}(\mathbf{x}_t, \mathbf{c}, t)\|^2}{2\sigma_t^2} \right) \right] \end{aligned} \quad (26)$$

其中  $\theta_{old}$  为旧策略参数，通过重要性采样比  $\frac{p_{\theta}}{p_{\theta_{old}}}$  实现策略的稳定更新。可进一步采用 PPO-2 clip 策略稳定训练。

**实现细节说明** 注：在理论推导中，L2 范数平方对应的损失函数采用总和归约（reduction='sum'），而实际代码实现中通常采用均值归约（reduction='mean'）以增强数值稳定性。这种理论推导与工程实践的差异性在 DDPM 的标准实现中同样存在，其主要差异体现在梯度幅值的缩放比例，而不会改变优化方向。具体表现为：

- 理论推导： $\|\cdot\|^2$  对应元素级平方和（sum）
- 工程实现： $\frac{1}{d}\|\cdot\|^2$  对应元素级平方均值（mean）

其中  $d$  为特征维度，这种缩放操作通过保持梯度量级在合理范围内，有效避免了训练过程中的数值溢出问题。以上推导中的  $\mu$  替换成 DDPM 中的  $\epsilon$ 、Rectified Flow 中的  $v$  和 score sde 中的 score 都成立，推导省略。

## 5 基于参考模型的扩散强化学习对齐方法 (DDPO with Reference Model)

$$\arg \max_{\theta} J = \underbrace{\mathbb{E}_{\mathbf{c} \sim D, x_0 \sim \pi_{\theta}(\cdot | c)} [r(x_0, c)]}_{\text{奖励最大化项}} - \underbrace{\beta \mathbb{E}_{\mathbf{c} \sim D, t \sim U(0, T)} [\mathbb{D}_{\text{KL}}(\pi_{\theta}(\cdot | x_t, c) \parallel \pi_{\text{ref}}(\cdot | x_t, c))]}_{\text{策略正则化项}} \quad (27)$$

根据 section 9 的推导结果，KL 散度的期望可以显式表达为：

$$\mathcal{L}_{DDPO_{KL}} = \mathbb{E}_{\mathbf{c} \sim D, t \sim U(0, T)} \left[ \frac{\|\boldsymbol{\mu}_{\theta}(x_t, c, t) - \boldsymbol{\mu}_{\text{ref}}(x_t, c, t)\|^2}{2\sigma_t^2} \right] \quad (28)$$

因此，完整的 DDPO 目标函数可分解为以下形式：

$$\begin{aligned}
\mathcal{L}_{DDPO_{total}} &= -(\mathcal{L}_{DDPO} - \beta \mathcal{L}_{DDPO_{KL}}) \\
&= -\mathbb{E}_{c \sim D, t \sim U(0, T)} \left[ r(x_0, c) \log \pi_{\theta}(x_{t-1} | x_t, c) \right] \\
&\quad + \beta \mathbb{E}_{c \sim D, t \sim U(0, T)} \left[ \frac{\|\boldsymbol{\mu}_{\theta}(x_t, c, t) - \boldsymbol{\mu}_{\text{ref}}(x_t, c, t)\|^2}{2\sigma_t^2} \right] \\
&= \mathbb{E}_{c \sim D, t \sim U(0, T)} \left[ \frac{r(x_0, c)}{2\sigma_t^2} \|x_{t-1} - \boldsymbol{\mu}_{\theta}(x_t, c, t)\|^2 \right. \\
&\quad \left. + \frac{\beta}{2\sigma_t^2} \|\boldsymbol{\mu}_{\theta}(x_t, c) - \boldsymbol{\mu}_{\text{ref}}(x_t, c, t)\|^2 \right]
\end{aligned} \tag{29}$$

## 6 Diffusion RLHF 中 DiffusionDPO 推导的谬误分析

若直接沿用 PPO 的奖励函数形式（如 eq. (8) 所示），将推导出错误的损失函数。其中  $\text{SG}()$  为梯度阻断函数，因其参数来自 RL 采样过程而非可微变量：

$$\begin{aligned}
\mathcal{L}_{DDPO_{wrong}} &= -\mathbb{E}_{c \sim D, t \sim U(0, T)} \left[ \text{SG} \left( r(x_0, c) - \beta \log \frac{\pi_{\theta}(x_{t-1} | x_t, c)}{\pi_{\text{ref}}(x_{t-1} | x_t, c)} \right) \right. \\
&\quad \left. \times \log \pi_{\theta}(x_{t-1} | x_t, c) \right] \\
&= \mathbb{E}_{c \sim D, t \sim U(0, T)} \left[ \text{SG} \left( r(x_0, c) - \beta \left( \frac{\|x_{t-1} - \boldsymbol{\mu}_{\text{ref}}(x_t, c, t)\|^2}{2\sigma_t^2} - \frac{\|x_{t-1} - \boldsymbol{\mu}_{\theta}(x_t, c, t)\|^2}{2\sigma_t^2} \right) \right) \right. \\
&\quad \left. \times \frac{\|x_{t-1} - \boldsymbol{\mu}_{\theta}(x_t, c, t)\|^2}{2\sigma_t^2} \right]
\end{aligned} \tag{30}$$

显然  $\mathcal{L}_{DDPO_{wrong}} \neq \mathcal{L}_{DDPO_{total}}$  成立。

特别考察  $\mathcal{L}_{DDPO_{wrong}}$  中的 KL 惩罚项：

$$\begin{aligned}
&= -\mathcal{L}_{DDPO_{wrong_{KL}}} \\
&= \frac{\beta}{4\sigma_t^4} \mathbb{E}_{c \sim D, t \sim U(0, T)} \left[ \text{SG} \left( \|x_{t-1} - \boldsymbol{\mu}_{\text{ref}}(x_t, c, t)\|^2 - \|x_{t-1} - \boldsymbol{\mu}_{\theta}(x_t, c, t)\|^2 \right) \right. \\
&\quad \left. \times \|x_{t-1} - \boldsymbol{\mu}_{\theta}(x_t, c, t)\|^2 \right]
\end{aligned} \tag{31}$$

其梯度表达式为：

$$\begin{aligned}
& -\nabla_{\boldsymbol{\mu}_\theta(\mathbf{x}_t, \mathbf{c}, t)} \mathcal{L}_{DDPO_{wrong_{KL}}} \\
& = \frac{\beta}{4\sigma_t^4} \mathbb{E}_{\mathbf{c} \sim D, t \sim U(0, T)} \left[ SG \left( \|\mathbf{x}_{t-1} - \boldsymbol{\mu}_{ref}(\mathbf{x}_t, \mathbf{c}, t)\|^2 - \|\mathbf{x}_{t-1} - \boldsymbol{\mu}_\theta(\mathbf{x}_t, \mathbf{c}, t)\|^2 \right) \right. \\
& \quad \left. \cdot \nabla_{\boldsymbol{\mu}_\theta} \|\mathbf{x}_{t-1} - \boldsymbol{\mu}_\theta\|^2 \right] \\
& = \frac{\beta}{4\sigma_t^4} \mathbb{E} \left[ SG \left( \|\mathbf{x}_{t-1} - \boldsymbol{\mu}_{ref}\|^2 - \|\mathbf{x}_{t-1} - \boldsymbol{\mu}_\theta\|^2 \right) (\boldsymbol{\mu}_\theta - \mathbf{x}_{t-1}) \right]
\end{aligned} \tag{32}$$

构造反例：当  $\boldsymbol{\mu}_{ref} > \mathbf{x}_{t-1} > \boldsymbol{\mu}_\theta$  且  $\|\mathbf{x}_{t-1} - \boldsymbol{\mu}_{ref}\|^2 > \|\mathbf{x}_{t-1} - \boldsymbol{\mu}_\theta\|^2$  时，可得  $-\nabla_{\boldsymbol{\mu}_\theta} \mathcal{L}_{DDPO_{wrong_{KL}}} < 0$ 。这表明即使  $\boldsymbol{\mu}_{ref} > \boldsymbol{\mu}_\theta$ ， $\boldsymbol{\mu}_\theta$  仍会朝减小的方向更新，存在明显逻辑悖论。

需特别注意，如 section 2.5 所述，DPO 的理论基础 eq. (65) 依赖于 eq. (8) 的有效性。若扩散模型的 RLHF 场景中 eq. (8) 不成立，则基于此导出的扩散 DPO 公式 eq. (33)（引自 Diffusion-DPO）将丧失理论支撑：

$$\begin{aligned}
L(\theta) = & -\mathbb{E}_{(\mathbf{x}_0^w, \mathbf{x}_0^l) \sim \mathcal{D}, t \sim \mathcal{U}(0, T), \mathbf{x}_t^w \sim q(\mathbf{x}_t^w | \mathbf{x}_0^w), \mathbf{x}_t^l \sim q(\mathbf{x}_t^l | \mathbf{x}_0^l)} \\
& \log \sigma \left( -\beta T \omega(\lambda_t) \left( \|\boldsymbol{\epsilon}^w - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t^w, t)\|_2^2 - \|\boldsymbol{\epsilon}^w - \boldsymbol{\epsilon}_{ref}(\mathbf{x}_t^w, t)\|_2^2 \right. \right. \\
& \quad \left. \left. - \left[ \|\boldsymbol{\epsilon}^l - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t^l, t)\|_2^2 - \|\boldsymbol{\epsilon}^l - \boldsymbol{\epsilon}_{ref}(\mathbf{x}_t^l, t)\|_2^2 \right] \right) \right)
\end{aligned} \tag{33}$$

## 7 修正的 DiffusionDPO

原始 DPO 梯度：

$$\begin{aligned}
& \nabla_\theta \mathcal{L}_{DPO}(\pi_\theta; \pi_{ref}) = \\
& -\beta \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[ \underbrace{\sigma(\hat{r}_\theta(x, y_l) - \hat{r}_\theta(x, y_w))}_{\text{higher weight when reward estimate is wrong}} \left[ \underbrace{\nabla_\theta \log \pi(y_w | x)}_{\text{increase likelihood of } y_w} - \underbrace{\nabla_\theta \log \pi(y_l | x)}_{\text{decrease likelihood of } y_l} \right] \right],
\end{aligned} \tag{34}$$



修正的 DiffusionDPO loss:

$$\begin{aligned}
\mathcal{L}_{\text{DiffusionDPO}}(\boldsymbol{\mu}_\theta(\mathbf{x}_t, \mathbf{c}, t); \boldsymbol{\mu}_{ref}(\mathbf{x}_t, \mathbf{c}, t)) = \\
- \beta \mathbb{E}_{(\mathbf{c}, \mathbf{x}_t^w, \mathbf{x}_t^l) \sim \mathcal{D}} \left[ \underbrace{SG[\sigma(\hat{r}_\theta(\mathbf{c}, \mathbf{x}_t^l) - \hat{r}_\theta(\mathbf{c}, \mathbf{x}_t^w))]}_{\text{higher weight when reward estimate is wrong}} \underbrace{\left[ -\frac{\|\mathbf{x}_{t-1}^w - \boldsymbol{\mu}_\theta(\mathbf{x}_t^w, \mathbf{c}, t)\|^2}{2\sigma_t^2} \right]}_{\text{increase likelihood of } \mathbf{x}_t^w} \right. \\
\left. + \underbrace{\frac{\|\mathbf{x}_{t-1}^l - \boldsymbol{\mu}_\theta(\mathbf{x}_t^l, \mathbf{c}, t)\|^2}{2\sigma_t^2}}_{\text{decrease likelihood of } \mathbf{x}_t^l} \right], \\
\hat{r}_\theta(\mathbf{c}, \mathbf{x}_t) = \beta \frac{\|\boldsymbol{\mu}_\theta(\mathbf{x}_t, \mathbf{c}, t) - \boldsymbol{\mu}_{ref}(\mathbf{x}_t, \mathbf{c}, t)\|^2}{2\sigma_t^2}
\end{aligned} \tag{35}$$

## 8 分类模型 KL 散度梯度推导

### 8.1 KL 散度的定义

KL 散度衡量两个离散概率分布  $\pi_\theta$  和  $\pi_{ref}$  之间的差异，定义为：

$$\begin{aligned}
\text{KL}(\pi_\theta \| \pi_{ref}) &= \mathbb{E}_{x \sim D, a_t \sim \pi_\theta(a_t | s_t, x)} \left[ \log \frac{\pi_\theta(a_t | s_t, x)}{\pi_{ref}(a_t | s_t, x)} \right] \\
&= \mathbb{E}_{x \sim D} \sum_{a_t \in \mathcal{Y}} \pi_\theta(a_t | s_t, x) \log \frac{\pi_\theta(a_t | s_t, x)}{\pi_{ref}(a_t | s_t, x)},
\end{aligned} \tag{36}$$

其中  $\mathcal{Y}$  为离散类别空间。

### 8.2 KL 散度梯度推导步骤

**步骤 1：展开 KL 表达式** 直接写出离散求和形式：

$$\text{KL}(\pi_\theta(\cdot | s_t, x) \| \pi_{ref}(\cdot | s_t, x)) = \sum_{a_t} \pi_\theta(a_t | s_t, x) \log \frac{\pi_\theta(a_t | s_t, x)}{\pi_{ref}(a_t | s_t, x)}. \tag{37}$$

**步骤 2：应用梯度算子** 对  $\theta$  求梯度：

$$-\nabla_\theta \text{KL}(\pi_\theta(\cdot | s_t, x) \| \pi_{ref}(\cdot | s_t, x)) = -\nabla_\theta \sum_{a_t} \pi_\theta(a_t | s_t, x) \log \frac{\pi_\theta(a_t | s_t, x)}{\pi_{ref}(a_t | s_t, x)}. \tag{38}$$

**步骤 3: 交换求和与梯度** 由于求和项有限且  $\pi_\theta(y|x)$  光滑, 可交换求和与梯度:

$$-\nabla_\theta \text{KL}(\pi_\theta(\cdot|s_t, x) \parallel \pi_{ref}(\cdot|s_t, x)) = -\sum_{a_t} \nabla_\theta \left[ \pi_\theta(a_t|s_t, x) \log \frac{\pi_\theta(a_t|s_t, x)}{\pi_{ref}(a_t|s_t, x)} \right]. \quad (39)$$

**步骤 4: 乘积法则分解** 对每一项应用乘积法则:

$$\begin{aligned} -\nabla_\theta \left[ \pi_\theta(a_t|s_t, x) \log \frac{\pi_\theta(a_t|s_t, x)}{\pi_{ref}(a_t|s_t, x)} \right] &= -\underbrace{[\nabla_\theta \pi_\theta(a_t|s_t, x) \cdot \log \frac{\pi_\theta(a_t|s_t, x)}{\pi_{ref}(a_t|s_t, x)}]}_{\text{Term 1}} \\ &\quad + \underbrace{\pi_\theta(a_t|s_t, x) \cdot \nabla_\theta \log \frac{\pi_\theta(a_t|s_t, x)}{\pi_{ref}(a_t|s_t, x)}}_{\text{Term 2}}. \end{aligned} \quad (40)$$

**步骤 5: 简化 Term 2** 由于  $\pi_{ref}(y|x)$  与  $\theta$  无关, 有:

$$\nabla_\theta \log \frac{\pi_\theta(a_t|s_t, x)}{\pi_{ref}(a_t|s_t, x)} = \nabla_\theta \log \pi_\theta(a_t|s_t, x) = \frac{\nabla_\theta \pi_\theta(a_t|s_t, x)}{\pi_\theta(a_t|s_t, x)}. \quad (41)$$

因此 Term 2 简化为:

$$\pi_\theta(a_t|s_t, x) \cdot \frac{\nabla_\theta \pi_\theta(a_t|s_t, x)}{\pi_\theta(a_t|s_t, x)} = \nabla_\theta \pi_\theta(a_t|s_t, x). \quad (42)$$

**步骤 6: 合并两项** 将 Term 1 和 Term 2 相加:

$$\sum_{a_t} \left[ \nabla_\theta \pi_\theta(a_t|s_t, x) \cdot \log \frac{\pi_\theta(a_t|s_t, x)}{\pi_{ref}(a_t|s_t, x)} + \nabla_\theta \pi_\theta(a_t|s_t, x) \right]. \quad (43)$$

**步骤 7: 处理归一化条件** 由于  $\sum_{a_t} \pi_\theta(a_t|s_t, x) = 1$ , 其梯度为 0:

$$\sum_{a_t} \nabla_\theta \pi_\theta(a_t|s_t, x) = \nabla_\theta \sum_{a_t} \pi_\theta(a_t|s_t, x) = \nabla_\theta 1 = 0. \quad (44)$$

因此第二项求和为 0, 仅保留第一项:

$$-\nabla_\theta \text{KL}(\pi_\theta(\cdot|s_t, x) \parallel \pi_{ref}(\cdot|s_t, x)) = -\sum_y \nabla_\theta \pi_\theta(a_t|s_t, x) \cdot \log \frac{\pi_\theta(a_t|s_t, x)}{\pi_{ref}(a_t|s_t, x)}. \quad (45)$$

**步骤 8: 对数导数技巧** 利用  $\nabla_{\theta} \pi_{\theta}(a_t|s_t, x) = \pi_{\theta}(y|x) \nabla_{\theta} \log \pi_{\theta}(a_t|s_t, x)$ , 改写为:

$$-\nabla_{\theta} \text{KL}(\pi_{\theta}(\cdot|s_t, x) \parallel \pi_{ref}(\cdot|s_t, x)) = -\sum_{a_t} \pi_{\theta}(a_t|s_t, x) \nabla_{\theta} \log \pi_{\theta}(a_t|s_t, x) \cdot \log \frac{\pi_{\theta}(a_t|s_t, x)}{\pi_{ref}(a_t|s_t, x)}. \quad (46)$$

**步骤 9: 期望形式** 最终梯度可表示为期望:

$$\begin{aligned} -\nabla_{\theta} \text{KL}_t(\pi_{\theta} \parallel \pi_{ref}) &= -\mathbb{E}_{x \sim D, a_t \sim \pi_{\theta}(a_t|s_t, x)} \left[ \nabla_{\theta} \log \pi_{\theta}(a_t|s_t, x) \cdot \log \frac{\pi_{\theta}(a_t|s_t, x)}{\pi_{ref}(a_t|s_t, x)} \right] \\ &= \mathbb{E}_{x \sim D, a_t \sim \pi_{\theta}(a_t|s_t, x)} \left[ \nabla_{\theta} \log \pi_{\theta}(a_t|s_t, x) \cdot \log \frac{\pi_{ref}(a_t|s_t, x)}{\pi_{\theta}(a_t|s_t, x)} \right]. \end{aligned} \quad (47)$$

**损失函数形式** 根据梯度公式 eq. (47), 损失函数可推导为:

$$\begin{aligned} \mathcal{L}_{\text{KL}_t}(\theta, \text{ref}) &= \mathbb{E}_{x \sim D, a_t \sim \pi_{\theta}(a_t|s)} \left[ \log \pi_{\theta}(a_t|s_t, x) \cdot \text{SG} \left( \log \frac{\pi_{\theta}(a_t|s_t, x)}{\pi_{ref}(a_t|s_t, x)} \right) \right] \\ &= \mathbb{E}_{x \sim D} \mathbb{E}_{a_t \sim \pi_{\theta}(a_t|s_t, x)} \frac{1}{2} \left[ \log \frac{\pi_{\theta}(a_t|s_t, x)}{\pi_{ref}(a_t|s_t, x)} \right]^2 \end{aligned} \quad (48)$$

其中  $\text{SG}(\cdot)$  表示阻断梯度 (代码实现中对应 ‘detach()’ 函数)。

## 9 高维高斯分布的 KL 散度和梯度的推导

考虑两个  $k$  维各向同性高斯分布:  $P \sim \mathcal{N}(\boldsymbol{\mu}_{\theta}, \sigma_1^2 \mathbf{I})$ ,  $Q \sim \mathcal{N}(\boldsymbol{\mu}_{\text{ref}}, \sigma_2^2 \mathbf{I})$   
其概率密度函数分别为:

$$\begin{aligned} P(\mathbf{x}) &= (2\pi\sigma_1^2)^{-k/2} \exp \left( -\frac{\|\mathbf{x} - \boldsymbol{\mu}_{\theta}\|^2}{2\sigma_1^2} \right) \\ Q(\mathbf{x}) &= (2\pi\sigma_2^2)^{-k/2} \exp \left( -\frac{\|\mathbf{x} - \boldsymbol{\mu}_{\text{ref}}\|^2}{2\sigma_2^2} \right) \end{aligned} \quad (49)$$

### 9.1 KL 散度定义

KL 散度定义为:

$$D_{\text{KL}}(P \parallel Q) = \mathbb{E}_{\mathbf{x} \sim P} \left[ \ln \frac{P(\mathbf{x})}{Q(\mathbf{x})} \right] \quad (50)$$

将密度函数代入，展开对数比：

$$\ln \frac{P(\mathbf{x})}{Q(\mathbf{x})} = \frac{k}{2} \ln \frac{\sigma_2^2}{\sigma_1^2} + \left[ -\frac{\|\mathbf{x} - \boldsymbol{\mu}_\theta\|^2}{2\sigma_1^2} + \frac{\|\mathbf{x} - \boldsymbol{\mu}_{\text{ref}}\|^2}{2\sigma_2^2} \right] \quad (51)$$

## 9.2 逐项计算期望

将 KL 散度拆分为常数项与二次项之和：

$$D_{\text{KL}} = \underbrace{\frac{k}{2} \ln \frac{\sigma_2^2}{\sigma_1^2}}_{\text{常数项}} + \underbrace{\mathbb{E}_P \left[ -\frac{\|\mathbf{x} - \boldsymbol{\mu}_\theta\|^2}{2\sigma_1^2} + \frac{\|\mathbf{x} - \boldsymbol{\mu}_{\text{ref}}\|^2}{2\sigma_2^2} \right]}_{\text{二次项}} \quad (52)$$

计算  $\mathbb{E}_P[\|\mathbf{x} - \boldsymbol{\mu}_\theta\|^2]$  由于  $\mathbf{x} \sim P$ ，协方差矩阵为  $\sigma_1^2 \mathbf{I}$ ，故：

$$\mathbb{E}_P[\|\mathbf{x} - \boldsymbol{\mu}_\theta\|^2] = \text{tr}(\sigma_1^2 \mathbf{I}) = k\sigma_1^2 \quad (53)$$

计算  $\mathbb{E}_P[\|\mathbf{x} - \boldsymbol{\mu}_{\text{ref}}\|^2]$  展开二次型并取期望：

$$\begin{aligned} \|\mathbf{x} - \boldsymbol{\mu}_{\text{ref}}\|^2 &= \|\mathbf{x} - \boldsymbol{\mu}_\theta + \boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}}\|^2 \\ &= \|\mathbf{x} - \boldsymbol{\mu}_\theta\|^2 + 2(\mathbf{x} - \boldsymbol{\mu}_\theta)^\top (\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}}) + \|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}}\|^2 \end{aligned} \quad (54)$$

取期望后，交叉项因  $\mathbb{E}_P[\mathbf{x} - \boldsymbol{\mu}_\theta] = 0$  而消失：

$$\mathbb{E}_P[\|\mathbf{x} - \boldsymbol{\mu}_{\text{ref}}\|^2] = k\sigma_1^2 + \|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}}\|^2 \quad (55)$$

代入二次项

$$\begin{aligned} \mathbb{E}_P \left[ -\frac{\|\mathbf{x} - \boldsymbol{\mu}_\theta\|^2}{2\sigma_1^2} + \frac{\|\mathbf{x} - \boldsymbol{\mu}_{\text{ref}}\|^2}{2\sigma_2^2} \right] &= -\frac{k\sigma_1^2}{2\sigma_1^2} + \frac{k\sigma_1^2 + \|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}}\|^2}{2\sigma_2^2} \\ &= -\frac{k}{2} + \frac{k\sigma_1^2}{2\sigma_2^2} + \frac{\|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}}\|^2}{2\sigma_2^2} \end{aligned} \quad (56)$$

合并所有项 将常数项与二次项合并：

$$\begin{aligned} D_{\text{KL}} &= \frac{k}{2} \ln \frac{\sigma_2^2}{\sigma_1^2} - \frac{k}{2} + \frac{k\sigma_1^2}{2\sigma_2^2} + \frac{\|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}}\|^2}{2\sigma_2^2} \\ &= \frac{k}{2} \left[ 2 \ln \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2}{\sigma_2^2} - 1 \right] + \frac{\|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}}\|^2}{2\sigma_2^2} \end{aligned} \quad (57)$$

**最终公式** 整理后得到：

$$D_{\text{KL}}(P \parallel Q) = k \underbrace{\left( \ln \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2}{2\sigma_2^2} - \frac{1}{2} \right)}_{\text{方差项}} + \underbrace{\frac{\|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}}\|^2}{2\sigma_2^2}}_{\text{均值项}} \quad (58)$$

当  $\sigma = \sigma_1 = \sigma_2$  时，

$$D_{\text{KL}}(P \parallel Q) = \frac{\|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}}\|^2}{2\sigma^2} \quad (59)$$

### 9.3 对 $\boldsymbol{\mu}_\theta$ 的梯度推导

从 KL 散度公式中提取与  $\boldsymbol{\mu}_\theta$  相关的项：

$$D_{\text{KL}}(P \parallel Q) = \underbrace{\frac{\|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}}\|^2}{2\sigma_2^2}}_{\text{唯一与 } \boldsymbol{\mu}_\theta \text{ 相关的项}} + \text{其他与 } \boldsymbol{\mu}_\theta \text{ 无关的项}. \quad (60)$$

**步骤 1：展开二次型** 对  $\|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}}\|^2$  展开：

$$\|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}}\|^2 = (\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}})^\top (\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}}). \quad (61)$$

**步骤 2：计算关于  $\boldsymbol{\mu}_\theta$  的梯度** 利用二次型的梯度公式：

$$-\nabla_{\boldsymbol{\mu}_\theta} [(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}})^\top (\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}})] = -2(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}}). \quad (62)$$

**步骤 3：组合梯度分量** 将梯度结果代入 KL 散度表达式：

$$-\nabla_{\boldsymbol{\mu}_\theta} D_{\text{KL}} = -\frac{1}{2\sigma_2^2} \cdot 2(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}}) = -\frac{\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}}}{\sigma_2^2}. \quad (63)$$

**最终梯度表达式**

$$-\nabla_{\boldsymbol{\mu}_\theta} D_{\text{KL}}(P \parallel Q) = -\frac{\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}}}{\sigma_2^2} \quad (64)$$

## 10 DPO 推导

本节基于 KL 约束的奖励最大化目标，推导出可操作的直接偏好优化目标函数。首先建立基础优化问题：

$$\begin{aligned}
& \max_{\pi} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi} [r(x, y)] - \beta D_{\text{KL}}[\pi(y|x) \| \pi_{\text{ref}}(y|x)] \\
&= \max_{\pi} \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y \sim \pi(y|x)} \left[ r(x, y) - \beta \log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)} \right] \\
&= \min_{\pi} \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y \sim \pi(y|x)} \left[ \log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)} - \frac{1}{\beta} r(x, y) \right]
\end{aligned} \tag{65}$$

为求解该优化问题，引入配分函数  $Z(x)$  构造概率分布：

$$Z(x) = \sum_y \pi_{\text{ref}}(y|x) \exp \left( \frac{1}{\beta} r(x, y) \right) \tag{66}$$

定义新的参考分布  $\pi^*$  为：

$$\pi^*(y|x) = \frac{1}{Z(x)} \pi_{\text{ref}}(y|x) \exp \left( \frac{1}{\beta} r(x, y) \right) \tag{67}$$

将目标函数重构为：

$$\min_{\pi} \mathbb{E}_{x \sim \mathcal{D}} [D_{\text{KL}}(\pi(y|x) \| \pi^*(y|x)) - \log Z(x)] \tag{68}$$

由于  $Z(x)$  与策略  $\pi$  无关，优化目标简化为最小化 KL 散度项。根据 KL 散度的非负性，当  $\pi = \pi^*$  时取得全局最优解。

## 10.1 从奖励建模到偏好学习

实际应用中直接求解  $\pi^*$  存在两大障碍：1) 真实奖励函数  $r^*$  未知；2) 配分函数  $Z(x)$  的计算涉及全响应空间积分。为此，我们引入偏好学习框架。

采用 Bradley-Terry 模型，对于输入  $x$  和响应对  $(y_w, y_l)$ ，偏好概率建模为：

$$p^*(y_w \succ y_l | x) = \frac{\exp(r^*(x, y_w))}{\exp(r^*(x, y_w)) + \exp(r^*(x, y_l))} \tag{69}$$

关键突破在于建立奖励函数与最优策略的显式关联。由式 (2) 可得：

$$r(x, y) = \beta \log \frac{\pi^*(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x) \tag{70}$$

将奖励差表达式代入偏好概率模型：

$$p^*(y_w \succ y_l | x) = \frac{1}{1 + \exp \left( \beta \log \frac{\pi^*(y_l|x)}{\pi_{\text{ref}}(y_l|x)} - \beta \log \frac{\pi^*(y_w|x)}{\pi_{\text{ref}}(y_w|x)} \right)} \tag{71}$$

## 10.2 最终目标函数

通过极大似然估计，得到直接优化策略的参数化目标函数：

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[ \log \sigma \left( \beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right] \quad (72)$$

## 11 重要性采样

### 11.1 基本概念与动机

重要性采样 (Importance Sampling) 是强化学习中实现离策略 (off-policy) 学习的关键技术，其核心思想是通过引入行为策略 (behavior policy) 的采样分布来估计目标策略 (target policy) 的期望值。这一方法在策略优化中具有双重意义：

1. \*\* 样本复用 \*\*：允许利用历史策略生成的旧样本进行当前策略更新，显著提升数据利用率
2. \*\* 方差控制 \*\*：通过重要性权重修正新旧策略的概率分布偏差，维持无偏估计特性

### 11.2 策略梯度推导

考虑策略梯度基本形式：

$$\nabla_\theta J(\theta) = \mathbb{E}_{(s_t, a_t) \sim \pi_\theta} [A_\theta(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t)] \quad (73)$$

当转换为离策略更新时，需要引入重要性权重 (Importance Weight)

$$\rho_t = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta'}(a_t|s_t)} :$$

$$\begin{aligned} \nabla_\theta J(\theta) &= \mathbb{E}_{(s_t, a_t) \sim \pi_{\theta'}} \left[ \frac{\pi_\theta(s_t, a_t)}{\pi_{\theta'}(s_t, a_t)} A_\theta(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t) \right] \\ &= \mathbb{E}_{(s_t, a_t) \sim \pi_{\theta'}} \left[ \frac{\pi_\theta(a_t|s_t) \pi_\theta(s_t)}{\pi_{\theta'}(a_t|s_t) \pi_{\theta'}(s_t)} A_\theta(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t) \right] \\ &\approx \mathbb{E}_{(s_t, a_t) \sim \pi_{\theta'}} \left[ \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta'}(a_t|s_t)} A_{\theta'}(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t) \right] \end{aligned} \quad (74)$$

推导过程中包含两个重要近似：

- 状态分布抵消假设：  $\pi_{\theta}(s_t) \approx \pi_{\theta'}(s_t)$ ，在策略更新幅度较小时成立
- 优势函数近似：  $A_{\theta}(s_t, a_t) \approx A_{\theta'}(s_t, a_t)$ ，要求新旧策略差异可控

### 11.3 目标函数形式化

令  $\theta' = \theta_{\text{old}}$ ，得到离策略目标函数：

$$J(\theta) = \mathbb{E}_{(s_t, a_t) \sim \pi_{\theta_{\text{old}}}} \left[ \frac{\pi_{\theta}(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)} A_{\theta_{\text{old}}}(s_t, a_t) \right] \quad (75)$$

该目标函数具有以下特性：

1. **无偏性**：当  $\pi_{\theta}$  与  $\pi_{\theta_{\text{old}}}$  的支撑集相同时保持无偏估计
2. **方差敏感性**：重要性权重  $\rho_t$  的数值稳定性直接影响梯度质量
3. **策略约束**：需通过 KL 散度等度量限制  $\pi_{\theta}$  与  $\pi_{\theta_{\text{old}}}$  的差异